On the Learning Dynamics of Label-Noise Memorization in ReLU MLPs

Yannis Kaltampanidis [1] **Mykola Pechenizkiy** [1] **Hannah Pinson** [1]

Abstract

Understanding the mechanisms of memorization in neural networks remains an open and challenging problem. In this work, we study label-noise memorization in two-layer ReLU MLPs through the learning dynamics of the first-layer weights. Our analysis suggests that label noise initially attenuates first-layer magnitude evolution while largely preserving weight directions, before inducing competing magnitude and directional dynamics between clean and noisy samples. Experiments on MNIST further suggest that these competing dynamics can reach a dynamical equilibrium prior to memorization, indicating that memorization can emerge without significant distortion of the first-layer weights.

1. Introduction

Deep neural networks can fit noisy labels while still generalizing to unseen data (Zhang et al., 2016), a phenomenon related to benign overfitting (Bartlett et al., 2019) and, more broadly, memorization of noisy or long-tailed samples (Feldman, 2020). Understanding how neural networks memorize is important not only for assessing the implications of memorization on generalization performance, but also for data privacy and safety. Prior works have identified a two-stage learning behavior in which networks first generalize on clean patterns before memorizing noise (Arpit et al., 2017), with memorization often emerging in deeper layers (Baldock et al., 2021). Others suggest that memorization is confined to a small set of neurons distributed across layers (Maini et al., 2023). Several theoretical works have studied benign overfitting in linear settings (Bartlett et al., 2019) or smoothed-ReLU networks (Frei et al., 2022). However, understanding how memorization emerges through neuron-level learning dynamics in ReLU networks remains an open question. In this work, we study label-noise-induced memorization in two-layer ReLU MLPs through the evolution of the first-layer weights. Our analysis suggests that label noise initially attenuates first-layer magnitude evolution while largely preserving weight directions, before inducing competing magnitude and directional dynamics between clean and noisy samples. Experiments on MNIST further suggest that these competing dynamics can reach a dynamical equilibrium prior to memorization, indicating that memorization can emerge without significant distortion of the first-layer weights.

Our work differs from (Kou et al., 2023; Han et al., 2025), as we do not consider a convolutional structure to independently process signal and noise. An extended related work section can be found in Appendix A.1.

[1]Eindhoven University of Technology, The Netherlands. Correspondence to: Yannis Kaltampanidis <i.kaltampanidis@tue.nl>.

Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

2. Preliminaries

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \{0,1\}$ denote a binary classification dataset. We consider a fully connected neural network with a single hidden layer of width m:

$$h_{i\alpha} = \mathrm{ReLU}\left(\mathbf{w}_\alpha^{(1)\top}\mathbf{x}_i + b_\alpha^{(1)}\right), \quad \alpha = 1, \dots, m$$

$$z_i = \sum_{\alpha=1}^{m} w_\alpha^{(2)} h_{i\alpha} + b^{(2)}$$

$$p_i = \sigma(z_i)$$

where $\mathbf{w}_\alpha^{(1)} \in \mathbb{R}^d$ and $b_\alpha^{(1)} \in \mathbb{R}$ are the input weights and bias of neuron α, $w_\alpha^{(2)} \in \mathbb{R}$ is the weight connecting neuron α to the output neuron, and $b^{(2)} \in \mathbb{R}$ is the output bias. Here, $\sigma(\cdot)$ denotes the sigmoid function. The model with parameters θ is trained by minimizing the empirical binary cross-entropy loss $\mathcal{L}(\theta)$. To study learning dynamics, we consider gradient flow, the continuous-time limit of gradient descent, as the learning rate $\eta \to 0$ (Saxe et al., 2019).

2.1. Learning Dynamics of the Hidden Layer

Applying the chain rule to the first-layer parameters yields:

$$\frac{d\mathbf{w}_\alpha^{(1)}}{dt} = w_\alpha^{(2)} \frac{1}{N} \sum_{i=1}^{N} e_i \mathbf{1}_{\{h_{i\alpha}>0\}} \mathbf{x}_i, \tag{1}$$

$$\frac{db_\alpha^{(1)}}{dt} = w_\alpha^{(2)} \frac{1}{N} \sum_{i=1}^{N} e_i \mathbf{1}_{\{h_{i\alpha}>0\}}. \tag{2}$$

where $e_i(t) = y_i - p_i(t)$ denotes the prediction residual and $\mathbf{1}_{\{\cdot\}}$ denotes the ReLU gating function, restricting the sum to samples satisfying $h_{i\alpha} > 0$. This motivates the time-varying *effective dataset* $D_\alpha(t) = \{i : \mathbf{w}_\alpha^{(1)\top}(t)\mathbf{x}_i + b_\alpha^{(1)}(t) > 0\}$. Then Eq. (1) can be written as:

$$\frac{d\mathbf{w}_\alpha^{(1)}(t)}{dt} = w_\alpha^{(2)}(t)\, G_\alpha(t)$$

where $G_\alpha(t) = \frac{1}{N}\sum_{i\in D_\alpha(t)} e_i(t)\,\mathbf{x}_i$ defines the effective learning signal for neuron α, given by a weighted average over samples in its effective dataset. Finally, decomposing $\mathbf{w}_\alpha^{(1)}(t)$ into magnitude $r_\alpha(t)$ and direction $\mathbf{u}_\alpha(t)$ yields:

$$\frac{dr_\alpha(t)}{dt} = w_\alpha^{(2)}(t)\,\langle G_\alpha(t), \mathbf{u}_\alpha(t)\rangle, \tag{3}$$

$$\frac{d\mathbf{u}_\alpha(t)}{dt} = \frac{w_\alpha^{(2)}(t)}{r_\alpha(t)}\,\Pi^{\perp}_{\mathbf{u}_\alpha(t)} G_\alpha(t) \tag{4}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product, and $\Pi^{\perp}_{\mathbf{u}} v = v - \langle v, \mathbf{u}\rangle\mathbf{u}$ is the projection of v onto the subspace orthogonal to $\mathbf{u}$. We refer to $G_\alpha(t)$ as the target vector, since $\mathbf{u}_\alpha(t)$ converges ($\frac{d\mathbf{u}_\alpha}{dt} = 0$) when $\mathbf{u}_\alpha(t)$ aligns with $G_\alpha(t)$, given that $w_\alpha^{(2)}(t) \neq 0$.

3. Memorization Dynamics in MLPs

In this work we study how fitting label-noise affects the dynamics of the hidden layer weights $\mathbf{w}_\alpha^{(1)}(t)$. As shown in Eqs. (3), (4), the update direction is determined by the target vector $G_\alpha(t)$, while $w_\alpha^{(2)}(t)$ scales its magnitude. Thus, understanding how label noise alters the target vector $G_\alpha(t)$ is central to understanding the mechanisms of memorization in this setting.

3.1. Noise Model

We consider training on a corrupted dataset $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^N$, where the observed labels $\tilde{y}_i$ are generated from clean labels $y_i \in \{0,1\}$ via independent symmetric label noise:

$$\tilde{y}_i = \begin{cases} y_i, & \text{with probability } 1-\rho, \\ 1-y_i, & \text{with probability } \rho, \end{cases} \qquad 0 < \rho < \tfrac{1}{2}.$$

We consider a label noise-rate $\rho < 0.5$ to retain informative signal. Equivalently, let $m_i \sim \text{Bernoulli}(\rho)$ independently, such that $\tilde{y}_i = (1-m_i)y_i + m_i(1-y_i)$.

3.2. Early Learning Regime

We study the dynamics near initialization for a 2-layer MLP with weights $\mathbf{w}^{(1)}, \mathbf{w}^{(2)} \sim \mathcal{N}(0, \sigma_w^2 I)$, $\sigma_w \ll 1$, and zero-initialized biases. For small initialization scales, weight norms evolve much slower than directions, $\frac{dr_\alpha(t)}{dt} \ll \left\|\frac{d\mathbf{u}_\alpha(t)}{dt}\right\|$, a phenomenon known as silent alignment (Atanasov et al., 2021). We can therefore treat $\frac{w_\alpha^{(2)}(t)}{r_\alpha(t)}$ as constant during this phase. Furthermore, the sigmoid can be approximated by zeroth-order $p_i(t) \approx \frac{1}{2}$, since the weights remain small.

As a neuron's direction $\mathbf{u}_\alpha(t)$ evolves, its effective dataset may also change. However, (Pinson, 2026) shows that under similar conditions, it is rapidly attracted to a stable effective dataset, which it subsequently retains during early learning. Experiments further suggest that $D_\alpha(0)$ differs only slightly from the stable dataset $D_\alpha^{*,\text{early}}$ (Appendix A.5). We therefore approximate each neuron by a fixed effective dataset $D_\alpha^{*,\text{early}}$ from initialization. This still differs from a linear model, where all neurons share the same effective dataset, namely the full dataset.

Proposition 3.1 (Early-time learning signal)**.** *Under the early-learning assumptions, the target vector can be expressed as:*

$$G_\alpha(t) \approx \frac{(1-2\rho)}{2}\, S_\alpha \tag{5}$$

where

$$S_\alpha = \frac{1}{N}\sum_{i\in D_\alpha(0)} (2y_i - 1)\mathbf{x}_i. \tag{6}$$

Proof: See Appendix A.2.

The vector S_α corresponds to the class-separation direction within $D_\alpha^{*,\text{early}}$, capturing the difference between positive and negative samples. Substituting into Eqs. (3), (4) yields:

$$\frac{dr_\alpha(t)}{dt} \propto \frac{(1-2\rho)}{2}\,\langle S_\alpha, \mathbf{u}_\alpha(t)\rangle \tag{7}$$

$$\frac{d\mathbf{u}_\alpha(t)}{dt} \propto \frac{1}{r_\alpha(t)}\,\frac{(1-2\rho)}{2}\,\Pi^{\perp}_{\mathbf{u}_\alpha(t)} S_\alpha \tag{8}$$

Therefore, in this regime, label noise attenuates the expected update dynamics by a factor $(1-2\rho)$ relative to the clean-label case, without changing the weight trajectory.

3.3. Onset of Memorization

Motivated by empirical evidence that neural networks first fit clean patterns before memorizing noise (Zhang et al., 2016; Arpit et al., 2017; Stephenson et al., 2021), we analyze the onset of memorization, where clean samples are confidently classified while noisy samples are not yet fitted.

We consider a regime in which the logits are large, $|z_i(t)| \gg 1$, so that the sigmoid operates near saturation, and the effective datasets $D_\alpha(t)$ remain locally stable. We further assume that over this time window, each effective dataset $D_\alpha(t)$ is independent of label noise, so that noisy labels remain uniformly distributed within the effective datasets. In this regime, samples are confidently classified with respect to their clean labels, while noisy samples are misclassified.

Defining the functional margin with respect to the clean label as $\mu_i(t) := (2y_i - 1)z_i(t)$, the sigmoid in the saturated region can be approximated by a first order asymptotic

expansion as $p_i(t) \approx y_i + (1 - 2y_i)e^{-\mu_i(t)}$.

Proposition 3.2 (Learning signal at the onset of memorization)**.** *Under the onset of memorization assumptions, the target vector can be expressed as:*

$$G_\alpha(t) \approx -\rho\, S_\alpha(t) + R_\alpha(t), \tag{9}$$

where

$$R_\alpha(t) = \frac{1}{N} \sum_{i \in D_\alpha(t)} (2y_i - 1)e^{-\mu_i(t)}\mathbf{x}_i.$$

Proof: See Appendix A.3.

The learning signal consists of a noise-induced drift $-\rho S_\alpha(t)$ and a margin-dependent correction $R_\alpha(t)$. The drift term opposes the early-learning class-separation direction to fit noisy samples, therefore reducing margins, while $R_\alpha(t)$ is a margin-weighted average over samples with weights scaling as $e^{-\mu_i(t)}$, increasing the relative contribution of lower-margin samples. Substituting into Eqs. 3 and 4 yields:

$$\mathbb{E}\left[\frac{dr_\alpha(t)}{dt}\right] \propto -\rho\langle S_\alpha(t), \mathbf{u}_\alpha(t)\rangle + \langle R_\alpha(t), \mathbf{u}_\alpha(t)\rangle, \tag{10}$$

$$\mathbb{E}\left[\frac{d\mathbf{u}_\alpha(t)}{dt}\right] \propto \frac{-\rho\Pi^{\perp}_{\mathbf{u}_\alpha(t)} S_\alpha(t) + \Pi^{\perp}_{\mathbf{u}_\alpha(t)} R_\alpha(t)}{r_\alpha(t)}. \tag{11}$$

Thus, the dynamics are governed by a competition between drift and correction: the drift term $-\rho S_\alpha(t)$ reduces alignment with the class-separation direction, while $R_\alpha(t)$ provides a margin-dependent correction. This interaction may lead to small adjustments in both magnitude and direction, allowing the network to fit noisy samples without significantly disrupting the learned features. A simplified toy-model illustrating these competition dynamics is provided in Appendix A.4.

Remark. Proposition 3.2 corresponds to a simplified setting in which the effective datasets are primarily determined by the underlying clean structure, while label noise acts as an independent perturbation. This holds in a simpler setting where the model is first trained on clean data, and label noise is subsequently introduced.

4. Experiments

To validate the theoretical assumptions and the proposed framework, we consider a balanced binary subset of the MNIST dataset, denoted by $\mathcal{D}$, consisting of $N = 3000$ total samples of the digits 1 ($y = 1$) and 0 ($y = 0$).

We define $\mathcal{D}_c$ and $\mathcal{D}_n$ as the clean (unaltered labels) and noisy (flipped labels) versions of $\mathcal{D}$, respectively. We consider two identical MLPs of hidden dimension $m = 1000$, $f_c(\theta_c)$ and $f_n(\theta_n)$, trained on $\mathcal{D}_c$ and $\mathcal{D}_n$. Both models are initialized with the same parameters, with weights sampled from $\mathcal{N}(0, 10^{-3})$, and biases initialized to zero. The models are then trained via gradient descent for $300k$ epochs, while we evaluate the dynamics for $\rho \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$.

4.1. Early Learning Regime

As described in Prop. 3.1, in the early learning regime, noise scales the first-layer magnitude dynamics by a factor $(1 - 2\rho)$ without affecting the corresponding directional dynamics. To evaluate this regime, we train f_n with learning rate $\eta = 10^{-3}$ and f_c with $(1 - 2\rho)\eta$, and compare their directional (Fig. 1, top-left) and magnitude (top-right) updates. Aligned with Prop. 3.1, we observe that during early training epochs, noise has negligible effect on the first-layer directional dynamics, while the corresponding magnitude dynamics are attenuated by a factor $(1 - 2\rho)$ (equivalent to a clean model trained with a scaled learning rate).

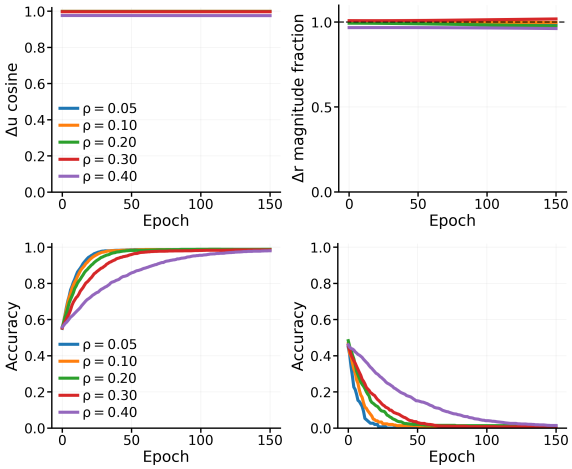


Figure 1. (top-left): cosine similarity between directional updates $\mathrm{sim}(\Delta_{u_c}, \Delta_{u_n})$; (top-right): radial-update norm ratio $\|\Delta r_n\|/\|\Delta r_c\|$; (bottom-left): accuracy of f_n on clean samples; (bottom-right): accuracy of f_n on noisy samples.

4.2. Onset of Memorization

High functional margin assumption. Proposition 3.2 assumes a training state in which all samples are confidently classified with respect to the clean (unflipped) labels, i.e., $\mu_i(t) \gg 1$ for all $i = 1, \ldots, N$. In this case, samples with corrupted labels are misclassified with respect to their observed (flipped) labels.

Fig. 2 illustrates the average functional margin $\tilde{\mu}_i(t)$, with respect to the **observed** labels, aggregated over clean (top-left) and noisy (top-right) samples. The corresponding model accuracies for clean and noisy samples are shown in the bottom row. We observe that as the noise rate increases, the deviation from the high-margin assumption is amplified, although the model continues to fit the underlying signal before adapting to corrupted labels.

Antagonistic dynamics of memorization. As suggested by Prop. 3.2, clean and noisy samples can induce competing effects on the first-layer weight magnitude $r_\alpha(t)$ and direction $\mathbf{u}_\alpha(t)$ dynamics. To evaluate this beyond the sim-

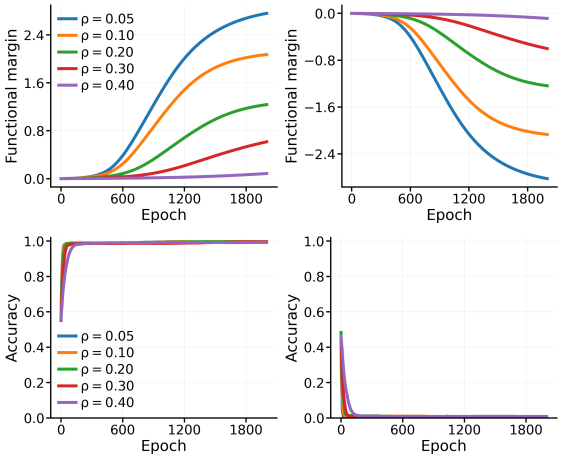


Figure 2. Functional margins (w.r.t observed labels) and accuracy of f_n prior to memorization. (top-left): clean-sample margins; (top-right): noisy-sample margins; (bottom-left): clean-sample accuracy; (bottom-right): noisy-sample accuracy.

plified theoretical setting, we analyze the training dynamics of f_n with noise ratio $\rho = 0.2$, aligned with experimental studies considering small or intermediate noise rates (Garg et al., 2023; Maini et al., 2023). We decompose the target vector as $G_\alpha(t) = G_{\alpha,c}(t) + G_{\alpha,n}(t)$, where $G_{\alpha,c}(t)$ and $G_{\alpha,n}(t)$ denote the clean and noisy contributions, respectively. As shown in Eqs. (3), (4), the inner products $\langle G_{\alpha,c}(t), \mathbf{u}_\alpha(t)\rangle$ and $\langle G_{\alpha,n}(t), \mathbf{u}_\alpha(t)\rangle$ govern the respective components to magnitude dynamics, while the orthogonal projections $\Pi^{\perp}_{\mathbf{u}_\alpha(t)} G_{\alpha,c}(t)$ and $\Pi^{\perp}_{\mathbf{u}_\alpha(t)} G_{\alpha,n}(t)$ drive directional evolution. By concatenating neuron-wise target vectors and directions into layer vectors G and $\mathbf{u}$, we study the collective first-layer dynamics.

Fig. 3 illustrates the competition dynamics between clean and noisy samples, on first-layer weights magnitude and direction. Throughout training, the directional components induced by clean and noisy samples remain negatively aligned ($\cos(\Pi^{\perp}_U G_c(t), \Pi^{\perp}_U G_n(t)) < -0.98$, see Appendix A.5), indicating that clean and noisy samples consistently attempt to rotate the first-layer weights in opposing directions.

During early training, indicated by the blue region, the magnitude components exhibit opposite signs, while $\|\langle G_c(t), U(t)\rangle\| > \|\langle G_n(t), U(t)\rangle\|$. Furthermore, clean and noisy samples induce opposing directional contributions while $\|\Pi^{\perp}_U G_c(t)\| \gg \|\Pi^{\perp}_U G_n(t)\|$ (Appendix A.5), yielding $\cos(\Pi^{\perp}_U G(t), \Pi^{\perp}_U G_c(t)) \approx 1$. Together, these observations validate the early-learning regime in which noise attenuates the evolution of the first-layer weight magnitude while having limited effect on the directional dynamics.

As clean samples become correctly classified, the corresponding prediction residuals decrease, reducing the contribution of $G_c(t)$ to both the magnitude and directional dynamics. In the orange region, competition emerges as $\|\langle G_n(t), U(t)\rangle\| \gtrsim \|\langle G_c(t), U(t)\rangle\|$, leading noisy samples to increasingly influence the magnitude evolution until $\langle U(t), G(t)\rangle \to 0$. At the same time, the alignment between the net directional component and the corresponding clean directional contributions is progressively reduced, indicating that noisy samples also inflict a directional drift. Notably, this transition occurs prior to memorization, suggesting that these competition dynamics can emerge without significantly altering the learned first-layer features. Interestingly, these competition dynamics appear to reach a dynamical equilibrium from this point onwards, where the net directional component oscillates between the competing clean and noisy directional contributions, while the net radial component remains close to zero.

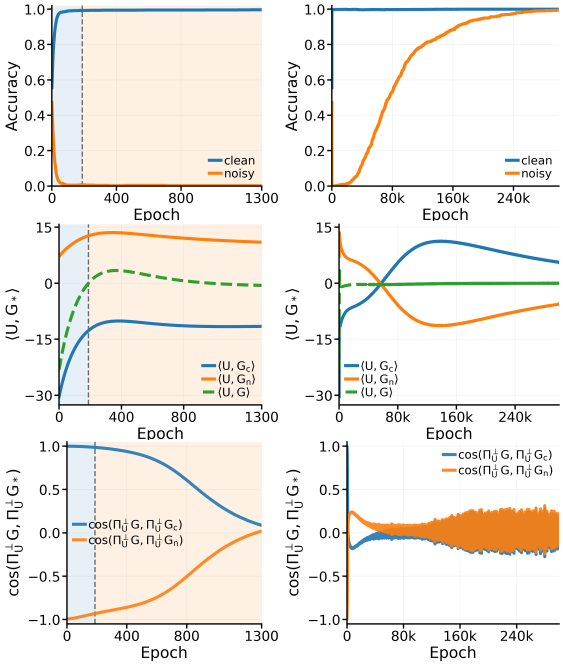


Figure 3. Early and full training dynamics of clean/noisy competition. Left column: first 1200 epochs. Right column: full training trajectory. (top): clean/noisy accuracy; (middle): magnitude-dynamics components $\langle G_c(t), U(t)\rangle$ and $\langle G_n(t), U(t)\rangle$; (bottom): alignment between the **net directional component** $\Pi^{\perp}_U G(t)$ and the corresponding clean/noisy directional components.

5. Conclusion

This work studies memorization under label noise in 2-layer ReLU MLPs through the learning dynamics of first-layer weights. Our analysis suggests that label noise induces competing directional and magnitude dynamics between clean and noisy samples, which can reach a dynamical equilibrium prior to memorization, indicating that memorization can emerge without dedicated "memorization neurons". This is a step towards understanding the mechanisms of memorization in standard ReLU networks, which is essential for assessing the privacy and safety risks associated with training sample memorization.

Acknowledgments

This work was carried out as a part of the ITEA 23020 MedGPT project.

References

Anagnostidis, S., Bachmann, G., Noci, L., and Hofmann, T. The curious case of benign memorization. *arXiv preprint arXiv:2210.14019*, 2022. URL https://arxiv.org/abs/2210.14019.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL https://proceedings.mlr.press/v70/arpit17a.html.

Atanasov, A., Bordelon, B., and Pehlevan, C. Neural networks as kernel learners: The silent alignment effect. *arXiv preprint arXiv:2111.00034*, 2021. URL https://arxiv.org/abs/2111.00034.

Baldock, R. J. N., Maennel, H., and Neyshabur, B. Deep learning through the lens of example difficulty. *arXiv preprint arXiv:2106.09647*, 2021. URL https://arxiv.org/abs/2106.09647.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2019.

Belkin, M., Hsu, D. J., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2:1167–1180, 2019.

Cao, Y., Gu, Q., and Belkin, M. Risk bounds for over-parameterized maximum margin classification on sub-Gaussian mixtures. *arXiv preprint arXiv:2104.13628*, 2021. URL https://arxiv.org/abs/2104.13628.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022. URL https://arxiv.org/abs/2202.06526.

Carlini, N., Erlingsson, Ú., and Papernot, N. Distribution density, tails, and outliers in machine learning: Metrics and applications. *arXiv preprint arXiv:1910.13427*, 2019. URL https://arxiv.org/abs/1910.13427.

Chatterji, N. S. and Long, P. M. Deep linear networks can benignly overfit when shallow ones do. *Journal of Machine Learning Research*, 24:117:1–117:39, 2022. URL https://jmlr.org/papers/v24/22-1276.html.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Frei, S., Chatterji, N. S., and Bartlett, P. L. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *arXiv preprint arXiv:2202.05928*, 2022. URL https://arxiv.org/abs/2202.05928.

Garg, I., Ravikumar, D., and Roy, K. Memorization through the lens of curvature of loss function around samples. *arXiv preprint arXiv:2307.05831*, 2023. URL https://arxiv.org/abs/2307.05831.

Han, A., Huang, W., Zhou, Z., Niu, G., Chen, W., Yan, J., Takeda, A., and Suzuki, T. On the role of label noise in the feature learning process. *arXiv preprint arXiv:2505.18909*, 2025. URL https://arxiv.org/abs/2505.18909.

Jiang, Z., Zhang, C., Talwar, K., and Mozer, M. C. Characterizing structural regularities of labeled data in overparameterized models. *arXiv preprint arXiv:2002.03206*, 2020. URL https://arxiv.org/abs/2002.03206.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign overfitting in two-layer ReLU convolutional neural networks. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/kou23a.html.

Li, Z., Zhou, Z.-H., and Gretton, A. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021. URL https://arxiv.org/abs/2106.03212.

Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151*, 2020. URL https://arxiv.org/abs/2007.00151.

Maennel, H., Alabdulmohsin, I. M., Tolstikhin, I. O., Baldock, R. J. N., Bousquet, O., Gelly, S., and Keysers, D. What do neural networks learn when trained with random labels? *arXiv preprint arXiv:2006.10455*, 2020. URL https://arxiv.org/abs/2006.10455.

Maini, P., Mozer, M. C., Sedghi, H., Lipton, Z. C., Kolter, J. Z., and Zhang, C. Can neural network memorization be localized? In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/maini23a.html.

Pinson, H. It's not a lottery, it's a race: Understanding how gradient descent adapts the network's capacity to the task, 2026. URL https://arxiv.org/abs/2602.04832.

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Stephenson, C., Padhy, S., Ganesh, A., Hui, Y., Tang, H., and Chung, S. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=V8jrrnwGbuc.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. URL https://arxiv.org/abs/1611.03530.

A. Appendix

A.1. Related Work

The statistical view of memorization as a synonym of overfitting was challenged by (Zhang et al., 2016), who showed that neural networks with sufficient capacity to achieve zero training loss can still generalize to unseen data, even without explicit regularization. (Arpit et al., 2017) further demonstrated that learning progresses in stages, with simple patterns fitted first, followed by harder and noisier samples. Based on these observations, (Feldman, 2020) formalized memorization as the *self-influence* of a training sample, defined by the change in its prediction when it is included versus excluded from the training set.

Memorization Proxy Metrics: Several works study memorization through metrics that correlate with atypical sample fitting. (Carlini et al., 2019) show that non-representative samples tend to exhibit lower *adversarial distance*. (Jiang et al., 2020) introduce the C-score, which measures the consistency of predictions between models trained on random subsets of the original dataset, finding that atypical samples exhibit low C-scores. More recently, (Garg et al., 2023) relate memorization to loss geometry, showing that the loss landscape exhibits high curvature around atypical samples.

Label noise induced memorization: Building on the observations of (Zhang et al., 2016; Arpit et al., 2017), another line of work studies memorization through the injection of label noise into the training data. In this setting, memorization is implicitly defined as the process of fitting noisy samples, providing an experimental framework that avoids reliance on computationally intensive self-influence metrics. (Baldock et al., 2021) introduce *prediction depth* as a measure of example difficulty, showing that early layers capture simple structure while deep layers memorize noisy samples. This is also supported by (Stephenson et al., 2021), who study the evolution of manifold geometric properties during training, showing that memorization emerges in deeper layers and later stages of training. However, (Maini et al., 2023) showcase that memorization is not confined to final layers, but can be localized to a small set of neurons distributed across the network. Finally, prior works (Maennel et al., 2020; Anagnostidis et al., 2022) study the extreme case of fully random labels, showing that useful representations may still emerge even in the absence of signal under certain conditions.

Theoretical work: Several theoretical works sought to explain the capacity of overparameterized models to fit noisy samples while achieving low test error, a phenomenon termed *benign overfitting* (Bartlett et al., 2019). Benign overfitting has been studied in a variety of settings, including linear regimes (Belkin et al., 2019; Cao et al., 2021; Chatterji & Long, 2022), neural tangent kernel (NTK) regimes (Li et al., 2021), and neural networks with smooth activations (Cao et al., 2022; Frei et al., 2022). (Kou et al., 2023) further extends these results to neural networks with ReLU activations. Finally, the two-stage behavior of generalization prior to memorization has been theoretically studied in linear models (Liu et al., 2020) and 2-layer homogeneous ReLU networks (Kou et al., 2023; Han et al., 2025), under simplified data and architectural assumptions.

A.2. Proof of Proposition 3.1

Under the early-learning assumptions, logits remain small and the sigmoid output can be approximated by zeroth order as $p_i(t) \approx \frac{1}{2}$, giving residuals

$$\tilde{e}_i(t) = \tilde{y}_i - p_i(t) \approx \tilde{y}_i - \frac{1}{2}.$$

Under the label-noise model $\tilde{y}_i = (1 - m_i)y_i + m_i(1 - y_i)$, a short calculation gives

$$\tilde{y}_i - \frac{1}{2} = \frac{1}{2}(1 - 2m_i)(2y_i - 1).$$

Substituting into the target vector:

$$G_\alpha(t) \approx \frac{1}{2N} \sum_{i \in D_\alpha(0)} (1 - 2m_i)(2y_i - 1)\mathbf{x}_i.$$

We partition $D_\alpha(0)$ into clean and noisy subsets

$$D_{\alpha,c} = \{i \in D_\alpha(0) : m_i = 0\}, \qquad D_{\alpha,n} = \{i \in D_\alpha(0) : m_i = 1\},$$

so that

$$G_\alpha(t) \approx \frac{1}{2N} \left[\sum_{i \in D_{\alpha,c}} (2y_i - 1)\mathbf{x}_i - \sum_{i \in D_{\alpha,n}} (2y_i - 1)\mathbf{x}_i \right].$$

Since $m_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\rho)$ independently of $(\mathbf{x}_i, y_i)$, for a large enough dataset the flipped samples form an approximately representative ρ-fraction of $D_\alpha(0)$, and thus

$$\sum_{i \in D_{\alpha,c}} (2y_i - 1)\mathbf{x}_i \approx (1-\rho) \sum_{i \in D_\alpha(0)} (2y_i - 1)\mathbf{x}_i, \qquad \sum_{i \in D_{\alpha,n}} (2y_i - 1)\mathbf{x}_i \approx \rho \sum_{i \in D_\alpha(0)} (2y_i - 1)\mathbf{x}_i.$$

Substituting back:

$$G_\alpha(t) \approx \frac{(1-\rho) - \rho}{2N} \sum_{i \in D_\alpha(0)} (2y_i - 1)\mathbf{x}_i = \frac{1 - 2\rho}{2} S_\alpha,$$

where $S_\alpha = \frac{1}{N} \sum_{i \in D_\alpha(0)} (2y_i - 1)\mathbf{x}_i$ depends only on the clean samples. □

A.3. Proof of Proposition 3.2

Under the memorization regime, we assume a state where the logits are large, $|z_i(t)| \gg 1$, so that the sigmoid operates near saturation, and the effective datasets $D_\alpha(t)$ remain locally stable. In this regime, samples are confidently classified with respect to their clean labels, while noisy samples are misclassified.

In the saturation region, for $|z_i(t)| \gg 1$ the sigmoid admits a first-order series approximation. Recall that for $|x| < 1$, the geometric series gives

$$\frac{1}{1+x} = 1 - x + x^2 - \cdots \approx 1 - x,$$

to first order. For $z \gg 1$, we have $e^{-z} \ll 1$, so setting $x = e^{-z}$:

$$\sigma(z) = \frac{1}{1+e^{-z}} \approx 1 - e^{-z}.$$

For $z \ll -1$, we have $e^{z} \ll 1$, so setting $x = e^{z}$:

$$\sigma(z) = \frac{e^{z}}{1+e^{z}} \approx e^{z}.$$

Defining the functional margin with respect to the clean label as $\mu_i(t) := (2y_i - 1)z_i(t)$, note that under the high-margin assumption $\mu_i(t) \gg 1$ for all i, since samples are confidently classified with respect to their clean labels. For a clean sample with $y_i = 1$, we have $z_i(t) \gg 1$ and $\mu_i(t) = z_i(t)$, so the first case applies. For a noisy sample with $y_i = 0$ misclassified as positive, we have $z_i(t) \ll -1$ and $\mu_i(t) = -z_i(t)$, so the second case applies. Both cases therefore unify as

$$p_i(t) = \sigma(z_i(t)) \approx y_i + (1 - 2y_i)e^{-\mu_i(t)},$$

which can be verified directly: for $y_i = 1$ this gives $1 - e^{-\mu_i(t)}$, and for $y_i = 0$ this gives $e^{-\mu_i(t)} = e^{z_i(t)}$, recovering both cases.

The residual is therefore

$$\tilde{e}_i(t) = \tilde{y}_i - p_i(t) \approx \tilde{y}_i - y_i - (1 - 2y_i)e^{-\mu_i(t)}.$$

Using the noise model $\tilde{y}_i = (1 - m_i)y_i + m_i(1 - y_i)$, we have

$$\tilde{y}_i - y_i = m_i(1 - 2y_i),$$

which yields

$$\tilde{e}_i(t) \approx (1 - 2y_i)\left(m_i - e^{-\mu_i(t)}\right).$$

Substituting into the target vector:

$$G_\alpha(t) \approx \frac{1}{N} \sum_{i \in D_\alpha(t)} (1 - 2y_i)\left(m_i - e^{-\mu_i(t)}\right)\mathbf{x}_i.$$

Since $m_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\rho)$ independently of $(\mathbf{x}_i, y_i)$, for a large enough dataset we approximate the empirical average of m_i over $D_\alpha(t)$ by its mean ρ, giving

$$G_\alpha(t) \approx \frac{1}{N} \sum_{i \in D_\alpha(t)} (1 - 2y_i) \left(\rho - e^{-\mu_i(t)}\right) \mathbf{x}_i = -\frac{1}{N} \sum_{i \in D_\alpha(t)} (2y_i - 1) \left(\rho - e^{-\mu_i(t)}\right) \mathbf{x}_i.$$

Defining

$$S_\alpha(t) = \frac{1}{N} \sum_{i \in D_\alpha(t)} (2y_i - 1)\mathbf{x}_i, \qquad R_\alpha(t) = \frac{1}{N} \sum_{i \in D_\alpha(t)} (2y_i - 1)e^{-\mu_i(t)}\mathbf{x}_i,$$

yields

$$G_\alpha(t) \approx -\rho\, S_\alpha(t) + R_\alpha(t).$$

□

A.4. A Reduced Toy Model for the Onset of Memorization

As shown in the previous section, the expected learning signal decomposes into two components: a noise-induced drift $-\rho S_\alpha(t)$, where ρ denotes the label noise rate, and a margin-dependent correction $R_\alpha(t)$. The drift term opposes the class-separation direction, reducing margins, while the correction term assigns larger weight to low-margin samples. This suggests a competition between fitting noisy samples and preserving clean margins.

To capture this interaction, we introduce a reduced toy model describing the evolution of average margins. Let $m_c(t)$ and $m_n(t)$ denote the average margins of clean and noisy samples, respectively, both measured with respect to the observed labels. We consider a regime in which the network has already learned the clean class structure, so that

$$m_c(0) \gg 1, \qquad m_n(0) \ll 0.$$

To model the trade-off between fitting noisy labels and preserving clean margins, we introduce the ansatz

$$\dot{m}_c(t) = -k\, \dot{m}_n(t), \qquad k > 0,$$

which yields the constraint

$$m_c(t) + k\, m_n(t) = C,$$

where the constant C is determined by initial conditions.

We model the learning signal through the pressure function

$$P(t) = (1 - \rho)e^{-m_c(t)} + \rho e^{-m_n(t)},$$

where $\rho \in (0, 1)$ is the label noise rate. The two terms capture the exponentially decaying contribution of well-classified clean and noisy samples.

Substituting $m_c(t) = C - k\, m_n(t)$ gives

$$P(t) = (1 - \rho)e^{-C} e^{k m_n(t)} + \rho e^{-m_n(t)}.$$

We then consider gradient flow on $P(t)$:

$$\dot{m}_n(t) = -\frac{dP}{dm_n}(t),$$

which yields

$$\dot{m}_n(t) = \rho e^{-m_n(t)} - k(1 - \rho)e^{-C} e^{k m_n(t)}.$$

These reduced dynamics make the competition explicit. The first term increases $m_n(t)$ when noisy samples are misclassified, while the second term grows as $m_n(t)$ increases, reflecting the cost of reducing the clean margin. The system therefore balances two opposing effects: pressure to fit noisy labels and resistance from degrading clean-sample margins.

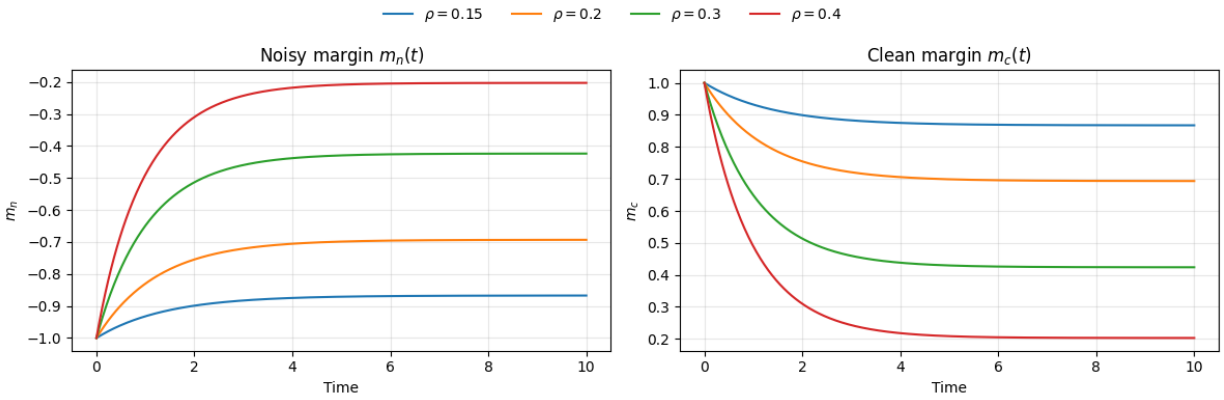


Figure 4. Dynamics of the reduced toy model. Noisy margins $m_n(t)$ increase while clean margins $m_c(t)$ decrease, with both trajectories converging to a stable trade-off. Higher noise levels ρ lead to stronger shifts toward fitting noisy samples.

Special case $k = 1$. Setting $k = 1$ and choosing symmetric initial conditions

$$m_n(0) = -1, \qquad m_c(0) = 1,$$

gives $C = 0$ and $m_c(t) = -m_n(t)$. The dynamics reduce to

$$\dot{m}_n(t) = \rho e^{-m_n(t)} - (1-\rho)e^{m_n(t)}.$$

Introducing $y(t) = e^{m_n(t)}$ yields

$$\dot{y}(t) = \rho - (1-\rho)y^2(t), \qquad y(0) = e^{-1},$$

which admits an explicit solution. Therefore,

$$m_n(t) = \log y(t), \qquad m_c(t) = -m_n(t).$$

As depicted in Fig. 4, this solution illustrates the qualitative behavior of the system: noisy margins increase from negative values, while clean margins decrease from positive values, and both converge to a stable trade-off determined by the noise level ρ.

A.5. Further Illustrations

Stable Effective Dataset Assumption. In Prop. 3.1 we approximate each neuron's effective dataset by its value at initialization, $D_\alpha(t) \approx D_\alpha(0)$, motivated by the result of (Pinson, 2026) that neurons are rapidly attracted to a stable effective dataset under similar conditions. To validate this empirically, we track how much the effective datasets of the noisy model drift relative to those of the clean model at initialization. For each neuron α, we define the binary gating vector $g_\alpha(t) \in \{0,1\}^N$, where

$$[g_\alpha(t)]_i = \mathbf{1}\left\{w_\alpha^{(1)\top}(t)\, x_i + b_\alpha^{(1)}(t) > 0\right\},$$

so that $[g_\alpha(t)]_i = 1$ if and only if sample $i \in D_\alpha(t)$. We measure the cosine similarity between the clean model's gating vector at initialization $g_\alpha^c(0)$ and the noisy model's gating vector at time t, $g_\alpha^n(t)$:

$$\cos(g_\alpha^c(0),\, g_\alpha^n(t)) = \frac{g_\alpha^c(0)^\top g_\alpha^n(t)}{\|g_\alpha^c(0)\|\,\|g_\alpha^n(t)\|},$$

and report the neuron-averaged quantity $\frac{1}{m}\sum_\alpha \cos(g_\alpha^c(0), g_\alpha^n(t))$ over the first 200 epochs, for $\rho \in \{0.05, 0.10, 0.20, 0.30, 0.40\}$.

As shown in Fig. 5, the cosine similarity remains high throughout early training across all noise rates, indicating that the effective datasets of the noisy model stay closely aligned with those of the clean model from initialization. This suggests

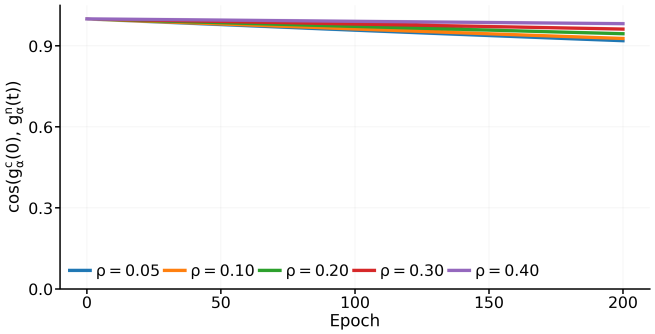
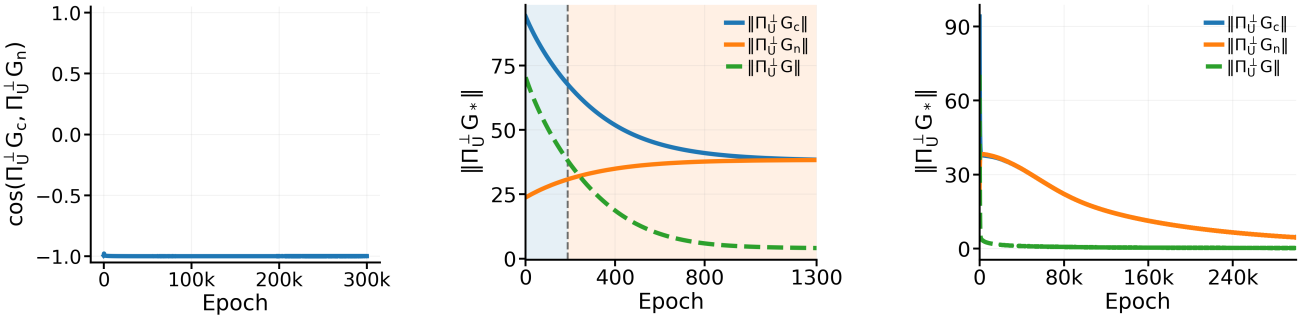


Figure 5. Neuron-averaged cosine similarity between the clean model gating vectors $g_\alpha^c(0)$ and the noisy model gating vectors $g_\alpha^n(t)$ over the first 200 epochs, for varying noise rates ρ. The cosine similarity remains close to 1 across all noise rates, supporting the stable effective dataset assumption of Prop. 3.1.

that label noise does not meaningfully distort the effective datasets during early learning, supporting the approximation $D_\alpha(t) \approx D_\alpha(0)$ used in Prop. 3.1.

Directional Competition Dynamics. We hereby provide additional experimental evidence to support the competition dynamics described in Section 4.2.



(a) *(b)* *(c)*

Figure 6. Directional competition dynamics. (a) Cosine similarity between the clean directional component $\Pi_U^\perp G_c(t)$ and the noisy contributions $\Pi_U^\perp G_n(t)$ over the full training trajectory. (b) Directional component norms during early training. (c) Directional component norms over the full trajectory.

Fig. 6(a) depicts the cosine similarity between the clean and noisy directional components $\Pi_U^\perp G_c(t)$ and $\Pi_U^\perp G_n(t)$ over the full training trajectory. The two components remain strongly antiparallel, with $\cos(\Pi_U^\perp G_c(t), \Pi_U^\perp G_n(t)) < -0.98$. During the early learning regime, $\|\Pi_U^\perp G_c(t)\| \gg \|\Pi_U^\perp G_n(t)\|$ (Fig. 6(b)), so the net directional component remains dominated by the clean contribution and no significant rotational drift is induced. As clean samples become correctly classified, $\|\Pi_U^\perp G_c(t)\|$ decays until $\|\Pi_U^\perp G_n(t)\|$ becomes comparable in magnitude (Fig. 6(c)), at which point the noisy component can induce a directional drift and the competition dynamics emerge.